05a: Questionnaire and Scale Development

1. Example of Scale Development Publication

Below are brief outlined steps in item development for establishing content validity. These steps were drawn from Holmbeck and Devine's (2009) checklist for scale development, and from the scale development work illustrated by Ragheb and Beard (1982) and Menon (2001).

An example for each step presented below will be drawn from a publication discussing the development and validation of an intercultural sensitivity scale. Below is the source with corresponding link.

> Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The intercultural development inventory. International journal of intercultural relations, 27(4), 421-443.
>
> http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-05a-Hammer-2003-Questionnaire-Development-Example.pdf

2. Preliminary Planning

Below are a few topics that must be considered first when developing, or selecting, questionnaires and scales. de Vaus (2002) discusses each of these in his Chapter 7 "Questionnaire Construction." A link to his chapter, found on the course web page under "Questionnaire/Scale Development" is also provided below.

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR8331-05a-de-Vaus-Chapter-7.pdf

Content

What should be included on questionnaires? For most research situations the study purpose and research questions or hypotheses will drive questionnaire content. Consider, for example, the following research questions:

(a) Is there a relation between math anxiety and math achievement?
(b) Does this relation differ between the sexes?
(c) Does the relation between anxiety and achievement vary by math complexity?
(d) Is the relation between math anxiety and achievement moderated by classroom autonomy support?

From these questions we know the following variables must be included in a questionnaire:

1. Scale to measure math anxiety
2. Measure of math achievement (either in questionnaire or as separate achievement measure, and if separate measure, then there must be a means to link student responses on the questionnaire to student responses on the achievement test, so this raises issues of anonymity and confidentiality that must be addressed)
3. Student sex
4. Math complexity – factual question like "Which course are you taking?" or "What is the most advanced math you have studied?" (algebra, trigonometry, calculus, etc.)
5. Scale to measure autonomy support

It is also important to collect contextual information about participants for descriptive purposes (e.g., sex, race, age, location, etc.) so study conditions, settings, and participants can be described.

While the example above fits within an educational setting, the logic for determining questionnaire content applies to most any setting – marketing research, health studies, etc.

Data Analysis Plans

It is critical to ensure you have collected data that is suitable for data analysis plans. Written responses to open-ended items won't work for most statistical modeling techniques, nor would math anxiety scale scores work for qualitative data analysis to learn why people fear math. When selecting or designing questionnaires, think carefully about how you plan to analyze collected data and whether the questionnaire items included will provide the types of scores suitable for data analysis planned.

Questionnaire Administration Plans

Self-administered questionnaires must be simple and easy to follow. Electronic self-administered questionnaires can have some built-in complexity and printed questionnaires should not. For example, it is easy to build electronic questionnaires with conditional or filter items, e.g., Do you shop online? If the answer is yes, respondents are directed to questions that ask about shopping behaviors and opinions; if the answer is no, respondents seamlessly skip the shopping behaviors and opinions items and move the next topic on the questionnaire. With paper questionnaires, filter items that instruct respondents to skip sections can lead to confusion and lower response rates. Questionnaires that are administered by trained interviewers can also be more complex since the interviewer can skip sections nearly as easily as electronic questionnaires.

Item Content Type

de Vaus (2002), Chapter 7, explains that questionnaire items fall into several categories: attributes, attitudes, behaviors, beliefs, and knowledge. See de Vaus for more discussion and illustrations.

de Vaus (2002) explains that when measuring attitudes, or related constructs (e.g., job satisfaction, math self-efficacy), the scale should be able to assess direction (e.g., positive vs negative, favor vs oppose, strong vs weak, etc.), extremity (i.e., how far respondent lies toward extreme positive – all top scores on self-efficacy, or all low scores on job satisfaction), and perhaps intensity (i.e., strength of attitude or thought). It is possible to have respondents at extreme positions, but have low intensity, so extremity and intensively are not the same although they are likely correlated.

3. Need for Instrument

If developing a new scale, one should

- review existing instruments and use established instrument if available and suitable; or
- explain possible need for, and contribution of new instrument developed.

Example 1: Below Hammer et al. begin describing theoretical and empirical need for an intercultural sensitivity measure, and they define intercultural sensitivity (p 422).

1. Introduction

As we begin the next millennium, the importance of effective intercultural relations in both global and domestic contexts is well recognized (Brislin, Cushner, Cherie, & Yong, 1986; Hammer, 1989, 1999a; Kealey, 1989). As Bhawuk and Brislin (1992) suggested, "To be effective in another culture, people must be interested in other cultures, be sensitive enough to notice cultural differences, and then also be willing to modify their behavior as an indication of respect for the people of other cultures" (p. 416). We will use the term "intercultural sensitivity" to refer to the ability to discriminate and experience relevant cultural differences, and we will use the term "intercultural competence" to mean the ability to think and act in interculturally appropriate ways. We argue that greater intercultural sensitivity is associated with greater potential for exercising intercultural competence.

4. Item Development

4a. Content Validity

- Define constructs so it is clear what will be measured (e.g. reading self-efficacy is…)
    - Recall that a construct is a variable that is "constructed" from responses to multiple items or indicators. Indictors are questionnaire items used to form a construct; indicators provide an indication—a measure—of respondents' positions on that which is measured.
    - Example indicators of reading self-efficacy:
        - In general, how confident are you in your abilities in reading?
        - How confident are you that you will do well in reading this year?
        - How confident are you that you can learn to be a good reader?
- Describe theory, if available, of construct – this overlaps with construct dimensions below.
- Identify and define construct dimensions, provide indicators of dimensions, e.g., dissertation process anxiety:
    - Physiological over-arousal (or emotionality): somatic (body, not mind) signs of anxiety and may include headaches, stomach aches, nausea, diarrhea, excessive sweating, shortness of breath, light-headedness or fainting, rapid heartbeat, and dry mouth.
    - Psychological - Worry: maladaptive cognitions, dread, negative thoughts. Include here catastrophic expectations of gloom and doom, fear of failure, random negative thoughts, feelings of inadequacy, self-condemnation, negative self-talk, frustration, comparing oneself unfavorably to others.
    - Psychological – Impairment: poor concentration, 'going blank' or 'freezing,' confusion, poor organization. The inability to concentrate leads to impaired performance on tests.

Example 2: Hammer et al. provide detailed explanation of theory used in development of their intercultural sensitivity scale, and this section also identifies and defines six dimensions of intercultural sensitivity (p 423)

2. Developmental Model of Intercultural Sensitivity

The Developmental Model of Intercultural Sensitivity (DMIS) was created by Bennett (1986, 1993b) as an explanation of how people construe cultural difference. Using a grounded theory approach (e.g., Glaser & Strauss, 1967; Strauss & Corbin, 1990), Bennett applied concepts from cybernetic constructivism (cf. Von Foerster, 1984; Brown, 1972; Maturana & Varela, 1987) to his observations of intercultural adaptation and identified six orientations that people seem to move through in their acquisition of intercultural competence. The underlying assumption of the model is that as one's *experience of cultural difference* becomes more complex and sophisticated, one's potential competence in intercultural relations increases.

According to this constructivist view, experience does not occur simply by being in the vicinity of events when they occur. Rather, experience is a function of how one construes the events (Kelly, 1963). The more perceptual and conceptual discriminations that can be brought to bear on the event, the more complex will be the construction of the event, and thus the richer will be the experience. In the case of intercultural relations, the "event" is that of cultural difference. The extent to which the event of cultural difference will be experienced is a function of how complexly it can be construed.

The set of distinctions that is appropriate to a particular culture is referred to as a *cultural worldview*. Individuals who have received largely monocultural socialization normally have access only to their own cultural worldview, so they are unable to

- Specify need for non-construct variables, i.e., observable, single-item, and demographic (i.e., attribute or factual) variables that are included in scale or questionnaire for descriptive, research, or validity assessment purposes (e.g., sex included because research shows females tend to demonstrate greater levels of test anxiety)

- Each dimension of construct should have separate item pool, with enough items to measure the dimension adequately, this is sometimes call item **sampling validity** which refers to the extent to which each dimension is covered by relevant items (e.g., on a statistics test we may develop either a table of specifications or performance objectives to help us identify areas that should be assessed and the number of items needed for each area, for example, maybe 3 items needed for correlation – how to interpret Pearson r, how to obtain Pearson r, and when to use Pearson r; thus a minimum of three items need for sampling validity on a statistics test which includes the content area of correlation)
- Develop item pool for construct (items will form a summated rating scale or index) and for non-construct variables
    - Items should be appropriate for intended population (e.g., use pictures for poor readers like ☺ 😐 ☹)
    - Sources of items:
        - Theory, deduction or brainstorming
        - Research examples
        - Questionnaires
        - Expert feedback
        - Target population feedback
        - Researcher experience

Example 3: Hammer et al. discuss development of initial item pool for scale (p 426)

3. Phase 1: developing an initial (60-item) version of the IDI

Some empirical research has been undertaken focused on developing preliminary measures of DMIS concepts (Pederson, 1998; Tower, 1990). However, these instruments were not subjected to psychometric testing. Therefore, we undertook an effort to develop a measure of the identified DMIS orientations following scale construction guidelines (e.g., DeVellis, 1991). This effort consisted of two phases. In the first phase, a preliminary, 60-item version of the IDI was developed. Subsequent testing of this version by Paige, Jacobs-Cassuto, Yershova and DeJaeghere (1999) suggested specific directions in further development of the IDI. In the second phase, we completed further analysis that resulted in a revised, 50-item IDI that is presented in this paper.

Example 4: Hammer et al. explained that they developed items based upon the six dimensions, but to confirm these dimensions could be identified empirically, they also interviewed participants from a variety of cultures to ascertain whether the six dimensions could be observed among cultural descriptions from participants, they also used participants' statements to form scale items (p 426)

We were initially concerned that the empirical observations upon which the DMIS was based could be re-created in systematic ways. This concern was addressed by examining discourse of people from a variety of cultures in order to determine if observers could reliably categorize the discourse in ways identified in the DMIS theoretical framework. A qualitative interview guide was designed to elicit perceptions of a group of respondents concerning their experience with cultural differences. This interview guide included questions that focused generally on how people experience cultural differences.[1] A research team was assembled, trained in cross-cultural interviewing techniques, and introduced to the DMIS. Following

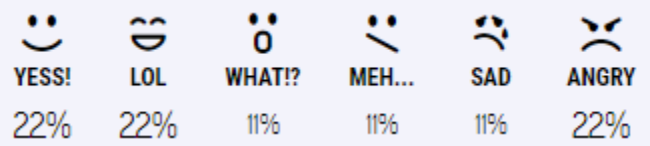
4b. Item Type: Closed-ended vs. Open-ended

- Types: Closed-ended/Structured/Forced-choice vs. Open-ended/Unstructured Item
- Closed-ended items have structured, predefined response options.
- Open-ended items allow respondents to provide responses in writing without reference to a list of possible responses.

Examples of closed-ended/structured items:

1. How did you learn about Georgia Southern?
   _____ From friends
   _____ Georgia Southern Website
   _____ Print ads
   _____ Radio ads
   _____ TV ads
   _____ etc.

2. What's your reaction?



Source: https://horrorfreaknews.com

3. Which of the following do you watch most?
   □ CNN
   □ Fox News
   □ MSNBC
   □ Other, please specify: _____

4. How would you rate your happiness with Amazon shopping?
   _____ Very Happy
   _____ Somewhat Happy
   _____ Somewhat Unhappy
   _____ Very Unhappy

Examples of Open-ended items:

1. Please explain why shop at Amazon instead competing online shopping sites.
2. How did you study for Test 2?
3. What did you think of the consolidation of Georgia Southern and Armstrong State University?

- Closed-ended items are
  - easier to code and score,
  - quicker to answer,
  - enable those who are less interested in writing to answer items easily;
  - and are more difficult and time consuming to develop,
  - may require many response options and therefore require careful thought and review (e.g., sex and gender classification; categories of birthday gifts; ways to study),
  - could force respondents to take positions or not answer items if their position is not included as a response option
- Open-ended items are
  - very good for exploratory research in which respondents' answers can be mined for detail and information,
  - better for obtaining details that cannot be obtained from closed-ended/structured items
  - are easier to write because one does not have to consider all possible answers before collecting data,
  - require much more time to code and analyze responses after data collection; are not good for respondents who are uninterested in writing.
- See Foliz (1996, p. 81) and de Vaus (2002, p. 99) discussion of differences between, and benefits of, open-ended and closed-ended items. Foliz's material is linked below and on the course web page.

Folz, D. H. (1996). Survey research for public administration. Sage Publications. Chapter 4 Survey Design and Implementation.

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-05a-Folz-1996-Chapter-4-Survey-Design.pdf

4c. Closed-ended Items: Scaling Methods

- Scales to measure attitudes and related constructs can be developed using a variety of scaling methods, but the most common is Likert.
- Oskamp and Schultz (2005), chapter 3, provide a good introduction and description of Thurstone, Guttman, Semantic differential, and Likert scaling formats. Their chapter can be found on the course web page and is linked below. A summary of these response options is presented below.

Oskamp, S., & Schultz, P. W. (2005). Attitudes and opinions (3rd edition). Psychology Press, Chapter 3 Explicit Measures of Attitude

http://www.bwgriffin.com/gsu/courses/edur8331/edur8331-presentations/EDUR-8331-05a-Oskamp-2005-Scaling-Methods-edited.pdf

- Likert-type scale – summated rating scale such as 1 = Very Dissatisfied to 7 = Very Satisfied)
    - For each scaled construct include one global, overall item to serve as content validity indicator, and empirical validity check via inter-item correlations; e.g.,
        - Latent variable = dissertation process self-efficacy: “Overall I am confident I can complete the dissertation successfully”
        - Latent variable = life satisfaction: “In general I am satisfied with my life.” This item can serve as construct validation for item analysis
- Osgood’s semantic differential – ask respondents to rate something on a variety of bipolar adjectives, with the desire to obtain ratings which indicate Evaluation, Potency, and Activity; for example, one may be asked to rate instructors:
    - Please rate your teacher with the following descriptions
    - Good 3 2 1 0 1 2 3 Bad
    - Weak 3 2 1 0 1 2 3 Strong
    - Active 3 2 1 0 1 2 3 Passive
    - etc.
- I find the semantic differential to be of little use in most research situations since it does not produce scale scores that easily interpretable.
- Thurstone – complex process; many items rated by panel of 100 or more; ratings are from 1 to 11, from least to most positive or similar directions; mean [or median] for each item determined; low variability desired; items with equal distance means [or medians] selected to form 10 item scale)
- Guttman – items are sorted so agreement with one means agreement with all preceding statements; each progressive statement represents a hardening or sharpening of opinion or knowledge; e.g.,
    - [a] 2+2=
    - [b] 2x2=
    - [c] 2x__=6
    - [d] (2/6)^4=

- Guttman scales are deterministic, one can predict responses based upon total score; if we know you answered item [c] above, then we can predict that you also correctly answered [b] and [a]; this is also the logic that forms Rasch analysis and Item Response Theory
- Other Item Formats (examples are provided below)
    - Rankings (e.g., sort items from most to least important)
    - Checklist (e.g., Which have you used to travel to work, check all that apply: [a] car, [b] walk, etc.)
    - Multiple Choice – like checklist, but usually only one selection from among possible options allowed, e.g.,
        - what is your marital status (married, never married, widowed, etc.)
        - which is the correct response to 2+2? (0, 1, 2, 3, 4, 5, etc.)
        - which is your biological sex: [a] female, [b] male
    - Feeling Thermometer – graphic means for respondents to rating things
    - Horizontal or Vertical scales – like Semantic differential with opposite positions (rather than adjectives) anchoring each end of the scale

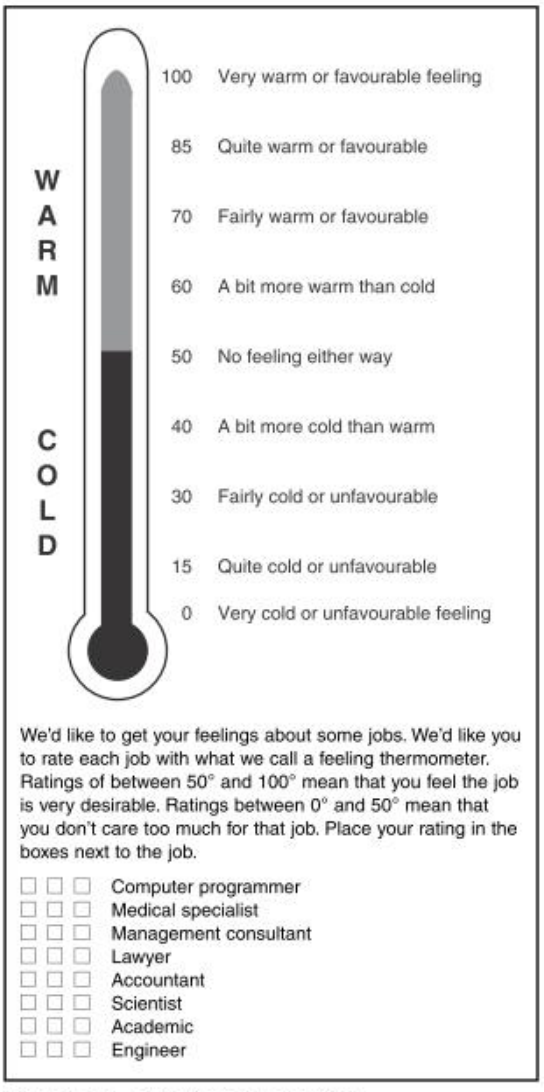


Figure 7.9 Feeling thermometer

Source: de Vaus (2002), p. 104

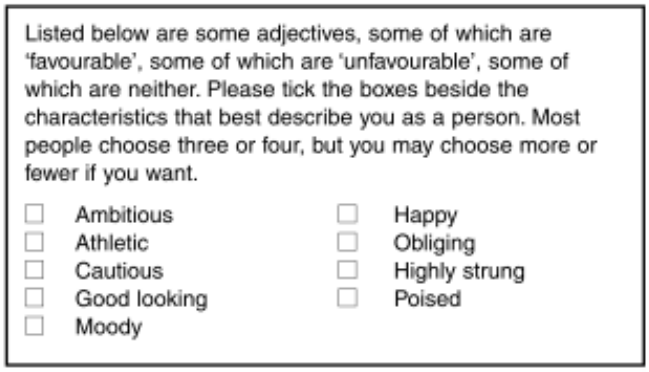


Figure 7.11 Checklist response format

Source: de Vaus (2002), p. 104

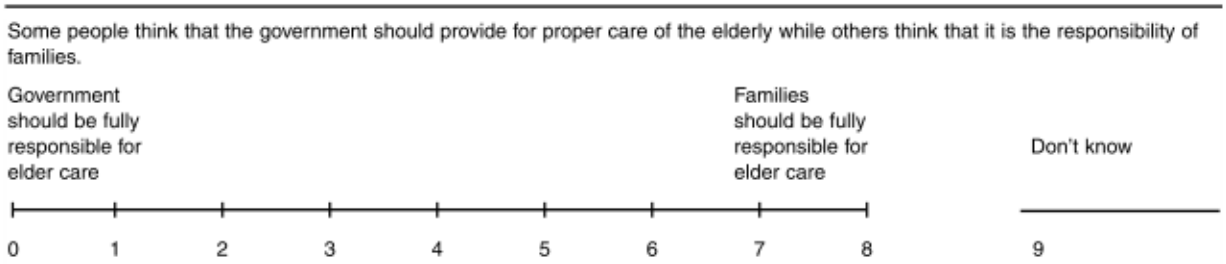


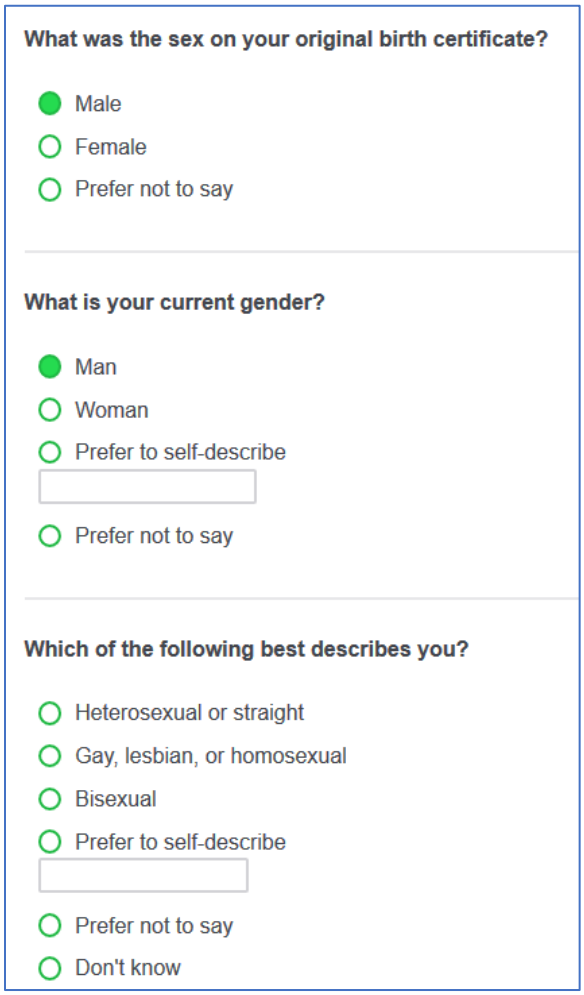
Figure 7.5 Horizontal rating scale

Source: de Vaus (2002), p. 102

4d. Closed-ended Items: Item Characteristics

- Discrimination – Good performing items discriminate among respondents; items should help us distinguish well the various positions respondents may hold; e.g., do our items help us measure differences among respondents with various levels (low, medium, high or even finer degradations) of
    - self-efficacy
    - life satisfaction
    - mathematics division skills
    - reading abilities, etc.
- de Vaus (2002) uses an example of income measurement –
    - poor discrimination item: What was your household income last year? (a) under 100,000 per year or (b) over 100,000 per year;
    - better discrimination item: What was your household income last year? (a) <30,000, (b) 30,000 to 60,000; (c) 60,001 to 90,000; (d) etc.
- Item Response Options
    - Exhaustiveness (or inclusiveness) – include all possible alternatives; this can be tedious and difficult for some topics, e.g., Which of the following represents ways you studied for Test 1?
        - Read book
        - Read notes
        - Worked examples in notes
        - Contacted instructor when I had questions
        - etc. – there can be many possible responses
    - Exclusiveness – responses enable one to fit in only one category (e.g., what is your race: Asian, Black/African-American, Latino, White) – sometimes not possible (e.g. what is your race: Asian-Black, Latino-Black, Asian-White, etc.) so multiple answers must be allowed; often such variables are measured with a checklist item where respondents may select more than one response; this can create coding and data analysis difficulties

Example of exhaustiveness difficulty; note how Gallup assesses sex and gender, but does not address transgender categories which would further expand response options required



Source: Gallup Poll, October 2018

- Non-committal responses – allow folks to answer "don't know" or "no opinion" when applicable to the item; forcing respondents to select a stance or opinion artificially creates responses that are not authentic and can be misleading for data analysis and reduce both reliability and validity of responses
- Number of response categories – for Likert scales, research suggest 5 or 7 seems optimal for maximizing reliability and validity; items with 3 response options (e.g., Agree, Disagree, Don't Know/Undecided) can work too for those items that don't fit well with a 5 or 7 response option, or if there are a large number of indicators that will be used to form the constructed variable (e.g., we will use 15 indicators with a 3-option response format to form the composite score of job satisfaction)
- Likert Response Options – most items with Likert-type scales follow a common Strongly Disagree to Strongly Agree response pattern. There are, however, many other options out there that make for more interesting items and offer variety. Below are links to two sources of some of these options – I encourage you to make use of these to expand the variety that is offered and to better hone items to fit your construct.
    - http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/04-questionnaire/04-likert-step-descriptions.htm
    - http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/04-questionnaire/04-LikertScaleOptions-Siegle.pdf
    - http://www.bwgriffin.com/gsu/courses/edur9131/2018spr-content/04-questionnaire/04-LikertScaleOptions-Vagias.pdf

  - The above are also linked on the course web page.
- Problematic responses
  - Social Desirability – respondents answer according to social norms, e.g., less likely to admit racist thoughts, more likely to indicate money donated to popular causes
  - Acquiescent Responses – agreeing with items regardless to item content
  - Response Set – tendency to mark items in the same way regardless of item content, e.g. answering all items with a score of 3 = neutral, neither disagree nor agree or with a score of 4 = agree.
- de Vaus (2002) offers some suggestions for addressing socially desirable answers – see p. 108

4e. Index vs Scale

- Index vs Scale (this is discussed in detail in "Reading Factor Analysis" notes)
  - Scale: Items should demonstrate internal consistency, be correlated, because response pattern on indicators should be correlated since the latent validity influences how people respond
  - Index: Items do not have to demonstrate internal consistency; sum of unrelated parts or items
    - SES – income, education, occupational prestige
    - Life Event Index – accumulation of milestones (e.g., high school graduation, college graduation, obtained career-oriented job, obtained promotion, etc.)

4f. Writing Items

It important that questionnaire items be clear, unambiguous, and a good fit for measuring the variable intended. de Vaus (2002) also offers useful tips on question wording starting on page 97 of his chapter on questionnaire development. The following figure is from de Vaus (2002, p. 97) and summarizes in question format his recommendations.

BOX 7.2 ***Question wording checklist***

1 Is the language simple?
2 Can the question be shortened?
3 Is the question double-barrelled?
4 Is the question leading?
5 Is the question negative?
6 Is the respondent likely to have the necessary knowledge?
7 Will the words have the same meaning for everyone?
8 Is there a prestige bias?
9 Is the question ambiguous?
10 Is the question too precise?
11 Is the frame of reference for the question sufficiently clear?
12 Does the question artificially create opinions?
13 Is personal or impersonal wording preferable?
14 Is the question wording unnecessarily detailed or objectionable?
15 Does the question have dangling alternatives?
16 Does the question contain gratuitous qualifiers?
17 Is the question a 'dead giveaway'?

Source: de Vaus (2002), p. 97

Below are suggestions for writing clear items based upon the work of Crocker and Algina (1986).

- Use as few items as possible to obtain valid scores; longer questionnaires reduce response rate
- Items should have one interpretation
    - Example sentence with at least two interpretations:
        - Poor: “You would be lucky to get him to work for you.”
        - Better: “His work performance is inconsistent and incomplete; rarely does he finish a task.”
- Avoid items to which most respondents agree or disagree because lack of variance in responses reduces item discrimination (the ability for an item to distinguish respondents on the concept measured)
    - Example
        - Most agree: “When people need help after facing devastating natural disasters, someone should be there to help them.”
        - Differences emerge: “When people need help after facing devastating natural disasters, a national government insurance policy supported by an income tax rate increase of 1.25% should be required.”
- Have a few items that are reversed to help prevent response set (marking items without carefully thinking about each item)
    - Example
        - I can learn the most challenging statistical analysis procedure taught in this class.
        - I believe I will perform well on statistical-related test items in this class.
        - The more difficult statistics become in this class, the less certain I am in learning those statistics. (Reversed response likely)
- Items should be as short as possible
- Avoid complex sentences with multiple segments, “if” or “because” links, etc.
    - Example
        - If $p \leq \alpha$ reject Ho, otherwise fail to reject Ho
- Items should be correct grammatically unless specific idiom or vernacular is intended.
- Items with absolute or indefinite qualifiers can create ambiguity or uncertainty of meaning
    - All, Always, None, Never
    - Only, Just, Merely, Many, Few, or Seldom
    - Example
        - Poor: “I am always washing my hands”
        - Better: “If possible, I wash my hands before eating”
- Use vocabulary that can be understood easily by respondents
    - Reading level checker can be helpful, e.g.
        - https://www.webpagefx.com/tools/read-able/
        - https://readable.io/
    - Example
        - Grade Level = 19:
            - “Some people have confidence in mathematics and some do not; statistics is based upon mathematics but also relies on logic and some folks have trouble with logic as well; how confident are you in your ability to learn complex statistics in this class?”
        - Grade Level = 10:
            - “How confident are you in your ability to learn complex statistics in this class?”
        - Grade Level = 6:
            - “Do you think you can learn statistics in this class?”
- Avoid use of negative (e.g., not, none, never)
    - Double negatives really create the problem, although negatives can as well. Ok to use negatives with some items, but always check for clarity.
    - Examples
        - Confusing: “I am not confident that I cannot learn statistics in this class.”

        - Response scale: Not true of me ---- Very true of me
    - Confusing: "I am not confident that I learn statistics in this class."
        - Response scale: Not true of me ---- Very true of me
    - Clearer: "I am confident that I can learn statistics in this class."
        - Response Scale: Not true of me ---- Very true of me
    - Clearer: "Learning statistics in this class is difficult for me."
        - Response Scale: Not true of me ---- Very true of me
- Items should focus on one construct (i.e., unidimensional); do not use double-barreled items
    - Examples of poor items
        - "Schools that perform poorly several years in a row should be closed and their teachers fired"
        - "I have the competence to work effectively and can influence the way work is done in my department"

5. Increasing Response Rate

- Short questionnaires – the number of items should be as few as possible because shorter questionnaires tend to have higher response and completion rates (Rolstad, Adler, & Rydén, 2011)
- Wording below are quotations from Edwards et al (2002) who performed a meta-analysis to identify factors related to response rates for postal surveys (and results apply for other surveys too); table on page 3 is particularly helpful.
    - "The odds of response were more than doubled when a monetary incentive was used (odds ratio 2.02; 95% confidence interval 1.79 to 2.27) and almost doubled when incentives were not conditional on response (1.71; 1.29 to 2.26).
    - Response was more likely when short questionnaires were used (1.86; 1.55 to 2.24).
    - Personalised questionnaires and letters increased response (1.16; 1.06 to 1.28),
    - as did the use of coloured ink (1.39; 1.16 to 1.67).
    - The odds of response were more than doubled when the questionnaires were sent by recorded delivery (2.21; 1.51 to 3.25)
    - and increased when stamped return envelopes were used (1.26; 1.13 to 1.41)
    - and questionnaires were sent by first class post (1.12; 1.02 to 1.23).
    - Contacting participants before sending questionnaires increased response (1.54; 1.24 to 1.92),
    - as did follow up contact (1.44; 1.22 to 1.70)
    - and providing non-respondents with a second copy of the questionnaire (1.41; 1.02 to 1.94).
    - Questionnaires designed to be of more interest to participants were more likely to be returned (2.44; 1.99 to 3.01),
    - but questionnaires containing questions of a sensitive nature were less likely to be returned (0.92; 0.87 to 0.98).
    - Questionnaires originating from universities were more likely to be returned than from other sources, such as commercial organisations (1.31; 1.11 to 1.54)."

6. Critical Item Analysis

Once developed, items should be critically reviewed. Below are suggestions for this process. Read each item carefully and assess the following:

- Wording clarity
- No redundancy within and across dimensions
- Fit with dimension and construct
- Fit with targeted population (e.g., is this item too complex for those to whom it will be administered)
- Fit with item scale - make sure the response scale logically fits with item wording, e.g.,
    - Poor fit: "Your level of satisfaction with current occupation"
        - 1= "Never or very infrequently" to 5 = "Almost every day or with high frequency"

    - Good fit: "Your level of satisfaction with current occupation"
        - 1 = "Very dissatisfied" to 5 = "Very satisfied".
- Remove or revise items as needed during this critical item analysis phase

7. Questionnaire Format

- See Fanning (2005) for many practical examples for formatting and layout of questionnaires.

  Fanning, E. (2005). Formatting a Paper-based Survey Questionnaire: Best Practices. Practical Assessment Research & Evaluation, 10. http://pareonline.net/pdf/v10n12.pdf

- Draft Questionnaire Format
    - Title
    - Brief introduction with general description of questionnaire purpose; be very general, not specific, since specificity could be leading (or misleading) and sway or bias responses
        - Example: If questionnaire measures four constructs – job satisfaction, job autonomy support, work-life conflict, and perceived work competence – instead of explaining these, just briefly explain that this questionnaire will ask you about the respondent's work experience.
    - Include instructions for completing and submitting questionnaire
    - If printed, best to use one side of paper, or be sure to include instructions at bottom of page ("See Back" or "Over please" or "Proceed to next page"). Questionnaires printed on both sides of page often result in missing data. The only exception to this is if the questionnaire is in booklet format.
- Develop and explain scoring plan for construct formation (Take mean of items 2, 3, 6, and 9 but use responses on 9 that are reversed scored)
- Leave adequate space for written answers
- Best to use vertical response options for multiple choice type questions because horizontal response options can be confusing about which answer space corresponds to the answer

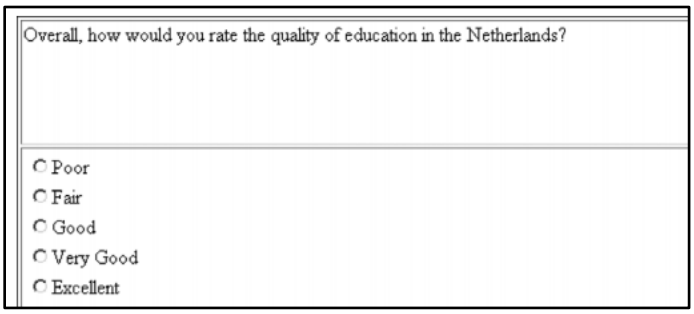
  Option A - Preferred
  Please indicate your race:
  _____ Asian
  _____ Black/African American
  _____ Latino
  _____ Mixed
  _____ White
  _____ Other

  Option B – Not Preferred, Leads to Confusion
  Please indicate your race:
  ____ Asian ____ Black/Afr. Amer. ____ Latino ____ Mixed ____ White ____ Other

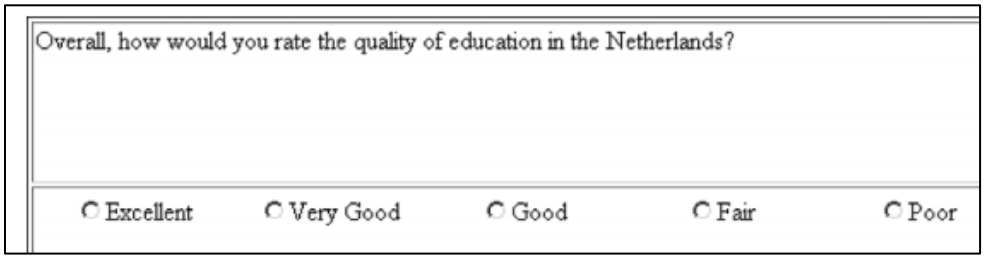
- For Likert-type items, I recommend horizontal formatting for quick responding by participants, and clear and easy data entry if using printed questionnaires – see below.

Course Evaluations – how would you rate…	Very Poor	Poor	Satisfactory	Good	Very Good
1. clarity of course material?	1	2	3	4	5
2. instructor responsiveness to students?	1	2	3	4	5
3. etc.	1	2	3	4	5

- Likert-type responses, should response options be positive to negative or negative to positive?
    - Negative to Positive: Poor, Fair, Good, Very Good, Excellent
    - Positive to Negative: Excellent, Very Good, Good, Fair, Poor
    - Friedman et al (1994) argue that direction can present a biasing effect in responses, however, their data in table 1 shows little evidence for this since only 3 of 10 items were significantly different.
    - Chan (1991) reached a similar conclusion to Friedman et al. but his findings were also mixed but generally more supportive of the bias.
    - Weng and Cheng (2000) found no evidence of bias from either order.
    - Overall research seems to be inconclusive about order bias; where there are differences, the differences do not seem large.
- Provide question instructions as needed, be brief (e.g., select all answers that apply; list only three items you bought recently)
- Question order
    1. Start with easy items for respondents to answer, maybe even enjoying answering; initial items should be clearly linked to purpose of study, if possible
    2. de Vaus (2002) states that one should not start questionnaire with demographic items (what is your age, sex, race); I disagree – no problem starting this way
    3. Order items from concrete to abstract (e.g., what is your race vs to what extent do you think your managers offer job autonomy to you?)
    4. Open-ended items should be few and near the end otherwise respondents won't complete questionnaire unless they are highly motivated for some reason (e.g., political questionnaire)
    5. With electronic or interviewer administered questionnaires, use filter questions to keep items relevant (e.g., if only females must answer questions 4 to 10, then automatically skip those items for males)
    6. Use both negatively and positively worded items with scales to break possible response set (e.g., I am confident I can work the most difficult math items on the upcoming test; I tend to find math tests difficult to answer)
    7. Try different question formats, if possible, to introduce variety; if using mostly Likert-type items, then word items so they use different response options (e.g., some items work with agree and disagree scales, others with frequency such as never, rarely, sometimes, etc., and maybe others with other adjectives (e.g. very poor to very good)
- Avoid contingency items with self-administered questionnaires because they create confusion (e.g., If you answer male to item 1 please skip to item 10, if you answered female to item 1 proceed to item 2, if you had an answer other than female or male, skip to item 16)
- Toepoel et al. (2009) conducted an experimental study of item layout with electronic questionnaires and found horizontal presentations seem to work better than vertical presentations, and linear better than non-linear.
    - Linear: Poor, Fair, Good, Very Good, Excellent
    - Non-linear (not on same line):
        - Poor, Fair, Good
        - Very Good, Excellent



vs.



Example 5: Hammer et al. developed many items, critically reviewed each, eliminated some and developed more, end result was initial draft of 239 items

> Once this base pool of items was identified, we reviewed the items for clarity, sentence structure, and ambiguous meanings. This resulted in the elimination of some of the items from the pool. Following this, we generated additional items for some of the dimensions based on our assessment of the number and comprehensiveness of the items identified in this initial pool (DeVellis, 1991). These additional items comprised less than 20% of the final list of 239 IDI sample items. Finally, the following response options were incorporated: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neutral, 5 = slightly agree, 6 = agree, 7 = strongly agree.

8. Expert Review

Knowledgeable individuations should critically review items for the same issues noted above in "Critical Item Analysis."

- Assess the following:
  - review definitions and dimensions of constructs
  - assess relevance of each item to construct
  - appropriateness of items and questionnaire for target population
  - reading level adequacy of items and questionnaire for target population
  - wording clarity
  - questionnaire format/layout
  - likelihood items may be objectionable to respondents
- Edit items and questionnaire as needed

Example 6: Following the pilot study, Hammer et al. asked a panel of experts to critically review the 239 items, and this review result in the item pool being reduced to 145 items

4. Panel review

A panel of experts then reviewed the item pool. This aided in further establishing the relevancy of the items to the construct of intercultural competence as well as providing initial reliability and validity estimates (DeVellis, 1991). The panel of experts was selected based on their demonstrated expertise within the intercultural field and familiarity with the DMIS in particular.[2]

A list of the 239 randomly ordered IDI items was sent to each expert to review. These individuals were asked to independently categorize the DMIS orientations or to check a response option of "unable to identify" if they felt the item could not be appropriately categorized. Inter-rater reliability among the expert ratings was then determined for each item. The criteria for selecting items from this analysis included the following: (1) a minimum of 5 of the 7 experts were able to categorize the item (i.e., if more than 2 of the 7 experts felt the item was too difficult to categorize, the item was eliminated from further consideration) and (2) inter-rater agreement for placing the item in the same category among the experts had to be 0.60 or above. These agreement criteria reflected exact agreement among 3, 4, or 5 of 5 raters; 4, 5, or 6 of 6 raters; and 5, 6, or 7 of 7 raters. Those items that could not either be categorized or reliably categorized by achieving an agreement rating of 0.60 or higher were eliminated.[3] Finally, each expert provided comments concerning each

9. Pilot Study (Field Test)

de Vaus (2002) suggests a three-part pilot test, briefly described below.

a. Partial pilot testing - Have experts evaluate items for fit and wording, pilot test items with respondents to ensure all interpret items in the same way, inform respondents that their participation is designed to seek improvement and ask respondents to mark confusing and poorly worded items, and provide feedback for improvement. de Vaus (2002, p. 116) offers the following guidance:
   - variation in item responses should be evident to help with item discrimination; little variation in responses to an item typically means it won't help distinguish respondents on the variable it was designed to measure
   - meaning of items should be clear to respondents – any confusion reduces validity and reliability
   - redundant items should be eliminated
   - internal consistency of scaled items should be high
   - items with large non-responses should be revised or eliminated – something is causing participants not to respond, so that item is problematic
   - check for response set patterns and patterns of missing data

b. Full pilot testing – administer questionnaire to large sample and use their responses to evaluate items statistically for item fit, and reliability and validity evidence, and seek respondents' feedback afterwards for suggestions for improvement; check time it takes respondents to complete questionnaire since goal is to make questionnaire completable in a short time, respondents' hopefully show interest and attention to questionnaire, whether contingency (filter) items work as planned, and whether respondents thought the questionnaire worked well – easy to understand, logical, etc.

c. Revised pilot testing – revise items as needed given results of the full pilot test, and make further pilot testing if substantial revisions were made (seek respondents' feedback afterwards for suggestions)

- Additional suggestions for pilot study
  - Use sample of respondents who match target population
  - Include open-ended item at end of questionnaire soliciting critical review and suggestions for revisions
  - Use a large a pilot sample if possible because a larger size allows for
    - Item analysis

- Reliability assessment (test-retest, internal consistency, equivalent forms)
- Rater/Coder Agreement Assessment
- Validity assessment (predicted differences, correlations, etc.)

Example 7: Hammer et al. piloted test the 239 items and asked participants to comment on item clarity and response option fit

> A pilot version of the IDI was then constructed using the 239 sample items. Two pilot test administrations of the IDI with culturally diverse groups of people were completed in order to identify difficulties respondents may have had with such issues as clarity of instructions, item clarity, response option applicability, and overall amount of time taken to complete the instrument. Based upon feedback from respondents, the IDI was further revised in these format areas.

Example 8: Following expert review, another pilot study, which they called Sample Testing, was conducted to assess psychometric properties of the scale (i.e., assess reliability and validity evidence)

> **5. Sample testing**
>
> This 145-item version of the IDI was administered to a sample of 226 subjects along with selected demographic items. The sample size approached the sample requirement recommended by Nunnally (1978) of 300 respondents for scale testing. The sample of respondents came from all walks of life and was not primarily drawn from a college student population. Of the 226 respondents, 43% were men ($n = 97$) and 57% were women ($n = 127$). The ages ranged from the low teens to over 60 years of age. The majority of respondents were between the ages of 22–30 (45%), with 10% under 21 years of age, 18% between 31 and 40, 16% between 41 and 50, 5% between 51 and 60 and 0.5% over 60 years of age. Thirty respondents were high school graduates (13%), 90 were college graduates (40%), 70 had M.A. or equivalent graduate degrees (31%), and 13 had Ph.D. or equivalent degrees (5%).

10. Data Entry (to be added)

Sections below explained in other presentation notes.

11. Reliability Assessment

- Test-retest
- Internal Consistency
- Parallel-forms
- Rater Agreement

12. Item Analysis

- Difficulty
- Discrimination
- Correlation with total score
- Contribution to reliability

13. Validity – Structural Assessment

- Correlation Matrix
- Exploratory Factor analysis
- Confirmatory Factor Analysis

14. Validity – Construct Assessment

- Construct
    - Correlated with related constructs
    - Mean differences with known groups
    - Correlated with similar measures
- Convergent – related as expected
- Divergent – unrelated as expected

References

Chan, J.C. (1991) Response-order Effects in Likert-type scales. Educational and Psychological Measurement, 51, 531-540.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Edwards, P., Roberts, I., Clarke, M., DiGuiseppi, C., Pratap, S., Wentz, R., & Kwan, I. (2002). Increasing response rates to postal questionnaires: systematic review. *BMJ : British Medical Journal*, *324*(7347), 1183.

Fanning, E. (2005). Formatting a Paper-based Survey Questionnaire: Best Practices. Practical Assessment Research & Evaluation, 10.

Friedman, H.H., Paul J. Herskovitz and Simcha Pollack (1994), "Biasing Effects of Scale-Checking Style in Response to a Likert Scale." Proceedings of the American Statistical Association Annual Conference: Survey Research Methods, 792-795.

Hammer, M. R., Bennett, M. J., & Wiseman, R. (2003). Measuring intercultural sensitivity: The intercultural development inventory. International journal of intercultural relations, 27(4), 421-443.

Holmbeck, G. N., & Devine, K. A. (2009). Editorial: An author's checklist for measure development and validation manuscripts. Journal of Pediatric Psychology, 1-6, 2009.

Menon, S.T. (2001). Employee empowerment: An integrative psychological approach. Applied Psychology: An International Review, 50, 153-180.

Ragheb, M. G., & Beard, J. G. (1982). Measuring leisure attitude. Journal of Leisure Research, 14, 155-167.

Rolstad, Adler, & Rydén (2011). Response Burden and Questionnaire Length: Is Shorter Better? A Review and Meta-analysis, Value in Health, 14, 1101-1108.

Toepoel V, Das M, van Soest A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales. Journal of Official Statistics, 25, 509–528.

Weng, L. & Cheng, CP (2000). Effects of Response Order on Likert-type Scales. Educational and Psychological Measurement, 60, 908-924.